NPS-AM-11-C8P10R03-044

EXCERPT FROM THE

PROCEEDINGS

OF THE

EIGHTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM WEDNESDAY SESSIONS VOLUME I

Utilizing Statistical Inference to Guide Expectations and Test Structuring During Operational Testing and Evaluation

Joy Brathwaite, Georgia Institute of Technology, Alton Wallace and Robert Holcomb, Institute for Defense Analyses

Published: 30 April 2011

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
APR 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Utilizing Statistical Inference to Guide Expectations and Test Structuring During Operational Testing and Evaluation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Georgia Institute of Technology,Atlanta,GA,30332**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Naval Postgraduate School?s 8th Annual Acquisition Research Symposium, 10-12 May 2011, Seaside, CA.

14. ABSTRACT

Comparative tests are commonly used during the operational testing phase to baseline the system under test (SUT) against the current status quo. Depending on the type of SUT, the comparative test may be costly and resource intensive. Thus any insights which may be gleaned about the potential results of the test beforehand may provide guidance on (1) the potential benefits of conducting the test and (2) the structuring of the test. This paper offers a statistical approach to understanding the type of results which may emerge during comparative testing of the SUT. Specifically, we utilize the concept of statistical inference to determine the needed performance difference between the SUT and the baseline system. If performance differences are statistically different, there may be useful information to be gained from conducting the test as is. Performance differences, which are not statistically different, may indicate that the test should be restructured or postponed. In either case, the relevant decision-maker is provided with information about the potential results of the test beforehand in order to make an informed decision. We illustrate the method of statistical inference on a system which improves situational awareness on the battlefield. We define a number of comparative metrics used to evaluate the operational effectiveness of the baseline system and the SUT. From the notional situational awareness system presented in this paper, we demonstrate the insights which may be gleaned and the implications for operational testing using statistical inference.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**43**	

The research presented at the symposium was supported by the Acquisition Chair of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request Defense Acquisition Research or to become a research sponsor, please contact:

NPS Acquisition Research Program
Attn: James B. Greene, RADM, USN, (Ret.)
Acquisition Chair
Graduate School of Business and Public Policy
Naval Postgraduate School
555 Dyer Road, Room 332
Monterey, CA 93943-5103
Tel: (831) 656-2092
Fax: (831) 656-2253
E-mail: jbgreene@nps.edu

Copies of the Acquisition Sponsored Research Reports may be printed from our website www.acquisitionresearch.net

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Preface & Acknowledgements

During his internship with the Graduate School of Business & Public Policy in June 2010, U.S. Air Force Academy Cadet Chase Lane surveyed the activities of the Naval Postgraduate School's Acquisition Research Program in its first seven years. The sheer volume of research products—almost 600 published papers (e.g., technical reports, journal articles, theses)—indicates the extent to which the depth and breadth of acquisition research has increased during these years. Over 300 authors contributed to these works, which means that the pool of those who have had significant intellectual engagement with acquisition issues has increased substantially. The broad range of research topics includes acquisition reform, defense industry, fielding, contracting, interoperability, organizational behavior, risk management, cost estimating, and many others. Approaches range from conceptual and exploratory studies to develop propositions about various aspects of acquisition, to applied and statistical analyses to test specific hypotheses. Methodologies include case studies, modeling, surveys, and experiments. On the whole, such findings make us both grateful for the ARP's progress to date, and hopeful that this progress in research will lead to substantive improvements in the DoD's acquisition outcomes.

As pragmatists, we of course recognize that such change can only occur to the extent that the potential knowledge wrapped up in these products is put to use and tested to determine its value. We take seriously the pernicious effects of the so-called "theory–practice" gap, which would separate the acquisition scholar from the acquisition practitioner, and relegate the scholar's work to mere academic "shelfware." Some design features of our program that we believe help avoid these effects include the following: connecting researchers with practitioners on specific projects; requiring researchers to brief sponsors on project findings as a condition of funding award; "pushing" potentially high-impact research reports (e.g., via overnight shipping) to selected practitioners and policy-makers; and most notably, sponsoring this symposium, which we craft intentionally as an opportunity for fruitful, lasting connections between scholars and practitioners.

A former Defense Acquisition Executive, responding to a comment that academic research was not generally useful in acquisition practice, opined, "That's not their [the academics'] problem—it's ours [the practitioners']. They can only perform research; it's up to us to use it." While we certainly agree with this sentiment, we also recognize that any research, however theoretical, must point to some termination in action; academics have a responsibility to make their work intelligible to practitioners. Thus we continue to seek projects that both comport with solid standards of scholarship, and address relevant acquisition issues. These years of experience have shown us the difficulty in attempting to balance these two objectives, but we are convinced that the attempt is absolutely essential if any real improvement is to be realized.

We gratefully acknowledge the ongoing support and leadership of our sponsors, whose foresight and vision have assured the continuing success of the Acquisition Research Program:

- Office of the Under Secretary of Defense (Acquisition, Technology & Logistics)
- Program Executive Officer SHIPS
- Commander, Naval Sea Systems Command
- Army Contracting Command, U.S. Army Materiel Command
- Program Manager, Airborne, Maritime and Fixed Station Joint Tactical Radio System

- Program Executive Officer Integrated Warfare Systems
- Office of the Assistant Secretary of the Air Force (Acquisition)
- Office of the Assistant Secretary of the Army (Acquisition, Logistics, & Technology)
- Deputy Assistant Secretary of the Navy (Acquisition & Logistics Management)
- Director, Strategic Systems Programs Office
- Deputy Director, Acquisition Career Management, US Army
- Defense Business Systems Acquisition Executive, Business Transformation Agency
- Office of Procurement and Assistance Management Headquarters, Department of Energy

We also thank the Naval Postgraduate School Foundation and acknowledge its generous contributions in support of this Symposium.

James B. Greene, Jr.
Rear Admiral, U.S. Navy (Ret.)

Keith F. Snider, PhD
Associate Professor

Panel 10 – New Testing Protocols for the Open Architecture Era

Wednesday, May 11, 2011	
3:30 p.m. – 5:00 p.m.	**Chair: Captain Brian Gannon, USN**, Program Manager, Naval Open Architecture, PEO IWS ***Modeling Complex System Testing: Characterizing Test Coverage to Improve Information Return*** Karl Pfeiffer, Valery Kanevsky, and Thomas Housel, NPS ***Test Reduction in Open Architecture via Dependency Analysis*** Valdis Berzins, Peter Lim, and Mohsen Ben Kahia, NPS ***Utilizing Statistical Inference to Guide Expectations and Test Structuring During Operational Testing and Evaluation*** Joy Brathwaite, Georgia Institute of Technology, Alton Wallace and Robert Holcomb, Institute for Defense Analyses

Captain Brian Gannon—CAPT Gannon was born in Chicago, Illinois and received a commission in 1985 through the Naval Reserve Officer Training Corps program at the Illinois Institute of Technology. His formal education includes a Bachelor of Science in Mechanical Engineering from the Illinois Institute of Technology, a Master of Science in Astronautical Engineering from the Naval Postgraduate School, and a Master of Business Administration from the University of Phoenix.

His service tours include Electronics Readiness Officer, ASW Officer and CIC Officer onboard *USS Gary* (FFG-51) from 1986 to 1989; Combat Systems Instructor at the Surface Warfare Officer's School in Coronado, CA, from 1989 to 1992; Student in the Space Systems Engineering curriculum at the Naval Postgraduate School from 1992 to 1994; Aegis Project Officer at the Port Hueneme Division, Naval Surface Warfare Center from 1994 to 1998; AEGIS LEAP Intercept (ALI) Project Officer in the Navy Theater Wide Program Office (PMS 452) from 1998 to 2002; TBMD Section Head in the Aegis Combat System Engineering Program Office (PMS 400B) from 2002 to 2003; Combat Systems Officer on the Fleet Maintenance staff for Commander, United States Pacific Fleet from 2003 to 2005; Technical Representative for Surface Naval Weapons (PEO IWS 3.0) and Aegis Ballistic Missile Defense (PD 452) portfolio of programs at Raytheon Missile Systems in Tucson, AZ.

CAPT Gannon assumed his present duties as Major Program Manager Future Combat Systems and Open Architecture (PEO IWS 7.0) in October 2008.

Captain Gannon's personal awards include the Meritorious Service Medal (four awards), Navy Commendation Medal and the Navy Achievement Medal in addition to various service awards. He is married to the former Jean Raup of Alexandria, VA. He has three children: Brittany (18), Timothy (15), and Christopher (13).

Utilizing Statistical Inference to Guide Expectations and Test Structuring During Operational Testing and Evaluation

Joy Brathwaite—Currently a fourth year PhD in Aerospace Engineering in the Space Systems Design Lab at Georgia Tech. Her general research interests are in the area of the Design and Acquisition of Military Assets. Specific domains of interests include the concept of value and its integration into the acquisition process, and the impact of investment strategies on needs identification and weapon system selection. Prior to entering the doctoral program, Joy received a BS in Aerospace Engineering and an MS in Economics from Georgia Tech, and spent approximately one year working as a research assistant at the Caribbean Development Bank. [joy.brathwaite@gatech.edu]

Alton Wallace—Member, Research Staff, Institute for Defense Analyses (IDA), Alexandria, VA. He has over 30 years of experience in operational testing (OT), and is credited with recognizing the difficulty of obtaining statistical significance in comparative ground combat tests that led to the development of this paper. He holds the BS, MS and PhD in mathematics respectively from NC A&T State University, Penn State University and the University of Maryland. Dr. Wallace is a former military officer and received the Bronze Star Medal for actions in Vietnam.

Robert Holcomb—Holcomb received his Bachelor of Science degree from the United States Military Academy in 1973 and was commissioned as a Second Lieutenant of Field Artillery in the United States Army. He received his Master of Science degree in Operations Research from the Naval Postgraduate School in 1982, and his doctorate in Information Technology from George Mason University in May 2011. Since retirement from military service in 1993, he has been a member of the research staff at the Institute for Defense Analyses (IDA). He was awarded IDA's Andrew J. Goodpaster Award for Excellence in Research in 2007. He is currently the head of the Land Warfare group within the Operational Evaluation Division of IDA.

Abstract

Comparative tests are commonly used during the operational testing phase to baseline the system under test (SUT) against the current status quo. Depending on the type of SUT, the comparative test may be costly and resource intensive. Thus, any insights which may be gleaned about the potential results of the test beforehand may provide guidance on (1) the potential benefits of conducting the test and (2) the structuring of the test. This paper offers a statistical approach to understanding the type of results which may emerge during comparative testing of the SUT. Specifically, we utilize the concept of statistical inference to determine the needed performance difference between the SUT and the baseline system. If performance differences are statistically different, there may be useful information to be gained from conducting the test as is. Performance differences, which are not statistically different, *may* indicate that the test should be restructured or postponed. In either case, the relevant decision-maker is provided with information about the potential results of the test beforehand in order to make an informed decision. We illustrate the method of statistical inference on a system which improves situational awareness on the battlefield. We define a number of comparative metrics used to evaluate the operational effectiveness of the baseline system and the SUT. From the notional situational awareness system presented in this paper, we demonstrate the insights which may be gleaned and the implications for operational testing using statistical inference.

Introduction

Comparative tests are used during the operational testing phase to baseline the system under test (SUT) against the current status quo. Depending on the type of SUT and test complexity, the comparative test may be costly to administer and challenging to repeat. Thus, any insights which may be gleaned about the potential results of the test beforehand may provide guidance on (1) the potential benefits of conducting the test and (2) the structuring of the test. From the relevant decision-maker's perspective (e.g., Office of the Under Secretary of Defense for Acquisition, Technology and Logistics [OUSD(AT&L)], Department of Defense [DoD]), knowledge about the potential outcome of the comparative test and implications for test structuring may lead to a more cost-effective test execution, providing maximal information about the SUT performance under operational conditions given resources expended.

This paper offers an applied statistical approach to understanding the type of results which may emerge during comparative testing of the SUT a priori. Statistical analysis is commonly used in the physical and social sciences to understand, quantify, and evaluate differences between treatment groups and control groups (Wooldridge, 2003). The statistical analysis employed to evaluate differences may range from a numerical or graphical description of observed differences using descriptive statistics to a more complex analysis in understanding the implications of pattern differences while accounting for randomness using inferential statistics. The specific statistical approach employed depends on the type of scientific inquiry being conducted and the data available. For this paper, we were concerned with understanding the performance difference of the SUT relative to the baseline system during the comparative test. In particular, we utilized the concept of statistical inference to determine the needed performance difference between the SUT and the baseline system for statistical significance. Next, we highlighted the implications of this analysis for test structuring. If performance differences are statistically different, useful information may be gained from conducting the test as is. Performance differences which are not statistically different *may* indicate that the test should be restructured or postponed. In either case, the relevant decision-maker is provided with information about the potential results of the test beforehand in order to make an informed decision.

To present the utilization of statistical inference in understanding the SUT performance a priori, this paper is divided into the following sections. The section titled Statistical Inference in Operational Testing and Evaluation presents an overview of the acquisition process of weapons systems and discusses the use of statistical inference in operational testing and evaluation. In the section titled Application of Statistical Inference in Guiding Operational Test Expectations, we illustrated the method of statistical inference on a system, which improves situational awareness on the battlefield. We defined a number of comparative metrics used to evaluate the operational effectiveness of the baseline system and the SUT. In the section titled Potential Outcomes and Analysis we delved a bit further into the analysis of the potential outcome of the comparative test. From the situational awareness system presented in this paper, we demonstrated the insights, which may be gleaned and the implications for operational testing using statistical inference. A few assumptions were made in evaluating the potential outcome of the comparative test. The section titled Sensitivity Analysis tests the robustness of the derived conclusions in the Potential Outcomes and Analysis section to changes in these assumptions. The section titled Conclusion completes this study. Although information about the actual system in this study has been masked, the data and analysis is representative of the actual system, and the implications and conclusions of this paper remained consistent with those derived from the original study

Statistical Inference in Operational Testing and Evaluation

The acquisition of a weapons system is traditionally divided into five phases with each phase having the requisite milestone. An acquisition program is required to meet the specific statutory and regulatory requirements dictated by the milestone to proceed to the next phase. The Milestone Decision Authority (MDA) holds the responsibility for determining whether the requirements of the milestone have been meet and the weapons program may proceed to the next phase. The phases of the acquisition process are shown in Figure 1.

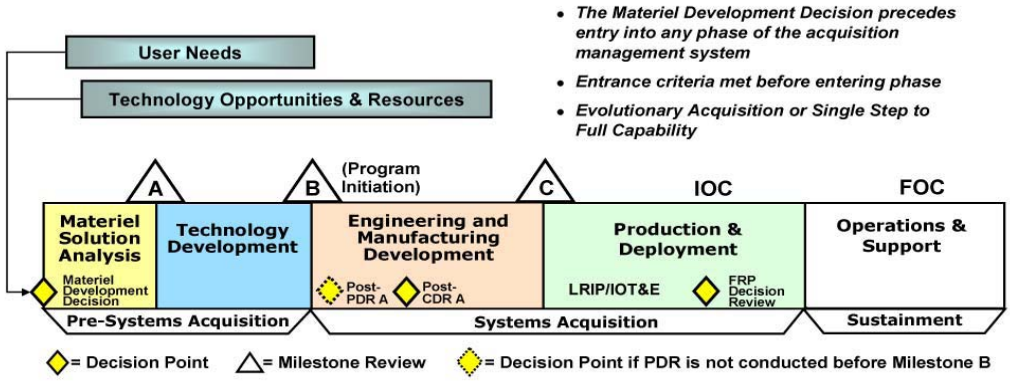


Figure 1. Overview of Acquisition of Weapon Systems

Note. Adapted from Department of Defense Instruction 5000.02 (USD[AT&L], 2008).

The first phase, the Materiel Solution Analysis, provides a preliminary analysis of the weapon systems. Within this phase, analysts first assess the user needs. If a need is shown to be evident, the analysts conduct an analysis of alternatives to evaluate probable options to fulfilling the need. The second phase, Technology Development, involves a determination of the technologies needed to operationalize the weapon system as well as the development and testing of the technology. Once the technology is shown to be functional in a relevant, or in the preferred case, an operational environment, the program may proceed to the third phase of the acquisition process. Within the third phase, Engineering and Manufacturing Deployment, the various sub-systems are developed, tested, and fully integrated into a complete weapon system. Also within this phase, a system demonstration occurs to show the military utility of the system as well as the manufacturing resources and processes required (Schwartz, 2010). The fourth stage is the Production and Deployment phase. In this phase, the weapon system enters Low Rate Initial Production (LRIP) and an Initial Operational Testing and Evaluation (IOT&E) occurs to determine the battle worthiness of the system. Congress requires testing of major systems and weapons programs to be conducted under operationally realistic conditions to determine the operational suitability of the system and whether it should proceed beyond LRIP (Fox, Boito, Graser, & Younossi, 2004). The final phase, Operations and Support involves a commitment to the full rate production and operation of the system. The system is fielded in a real time operational environment and maintenance support (among other types of support) provided by the relevant contractor(s).

During IOT&E, comparative evaluations are sometimes conducted. These tests are side-by-side comparisons in which the performance of a battalion with the SUT and without the SUT will be examined through a series of tactical battles. The intent of the test is to determine whether (and by how much) the unit's performance improves with the SUT. One method for assessing whether an improvement has occurred is through the use of statistical

inference. This technique, in particular significance testing, is a well-established method for determining whether the outcome of a treatment scenario differs significantly from a controlled scenario. Generally, statistical inference is used in the analysis of the outcome of operational tests and has been noted as a best practice in system evaluation (Commission on Behavioral and Social Sciences and Education [CBASSE], 1998). While it is not suggested that statistical inference is the sole evaluation tool, statistical inference possesses a number of advantages primarily among which is its objectivity given the various incentives and motivations of the stakeholders in the acquisition process. In addition, through statistical inference, it is possible to gain insight beforehand on the outcome of comparative tests.

In prior comparative tests, particularly for ground combat systems, there has been mixed success in establishing improved effectiveness of new systems using operational performance metrics. In most of these tests, a major contributor to the difficulty in finding a statistically significant "difference" between the performances of the unit with the SUT and without the SUT has been the sizable magnitude of the variability within the data for the metric being considered and the small sample size (CBASSE, 1995). That is, the standard deviation within the data has been so large that finding a difference between the means or other measures of central tendencies requires a really sizable difference in the means of the two data sets—generally larger than can be reasonably expected in combat.

The phenomenon is illustrated in Figure 2. Ideally, we expect to see what is shown on the left. The ideal data would show small performance variability by the SUT and the baseline system accompanied by a significant mean performance difference between the SUT and the baseline system. What frequently happens in ground combat tests is depicted on the right. The actual data commonly reveals large performance variability for both SUT and baseline systems as well as marginal differences in their mean performance.

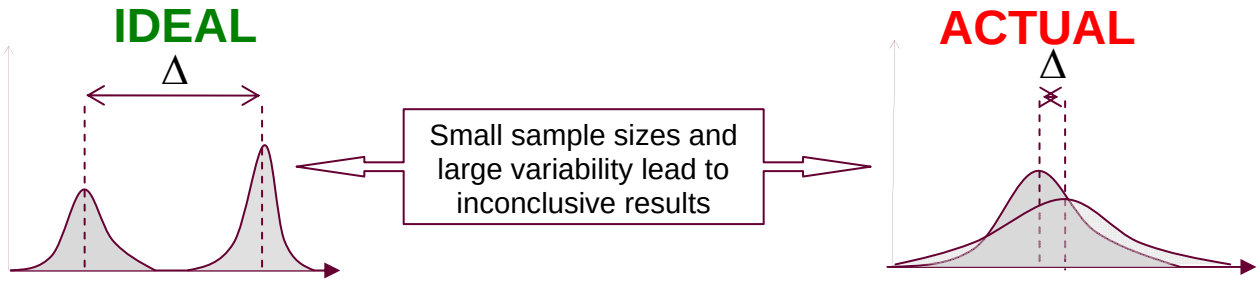


Figure 2. Illustration of Ideal Versus Actual Variability in Metrics

In the end, the data generally show that any "apparent" difference between the performance of the unit with and without the new capability has proven to be not statistically "significant." Upon finding "no difference" in operational performance metrics, alibis are offered for the technical results (e.g., unit was not trained, poor scenarios), and the assessments resort to subjective measures (e.g., interview comments, commanders impressions) to support the case for buying the new system.

In order to avoid such situations, we propose that it is possible to examine possible outcomes of the comparative tests beforehand through the application of statistical inference. The outcome of an a priori analysis using statistical inference may point to areas where the test may be modified or additional control measures may be introduced to increase the likelihood of obtaining desired results or highlighting scenarios in which utilization of the SUT may not be appropriate. In the next section, we demonstrate the use of

statistical inference on a system designed to improve situational awareness and describe how the results may be used to guide expectations on the outcome of the comparative tests.

Application of Statistical Inference in Guiding Operational Test Expectations

One of the key insights emerging from Operation Joint Endeavor and Operation Desert Storm was the need for a mobile infantry that would rapidly deploy ahead of the armed forces. To support this rapid deployment, a number of operational needs statements (ONS) from theater called for ground and aerial robotic capability to enable better situation awareness and understanding. These systems would improve intelligence, surveillance and reconnaissance at the lower unit level through advanced technological capabilities and allow speedy intelligence dissemination through enhanced networking capabilities. Over the last two decades, a number of systems have been or are being developed to address the issue of situational awareness and understanding. Examples of these include the Battlefield Combat Identification System (BCIS), a secure question and answer system that was intended to perform active identification of friendly targets to minimize fratricide on the battlefield, and the Early Infantry Brigade Combat Team systems (nee Future Combat System), which were intended to rapidly and securely disseminate information, thereby providing a technological advantage over the enemy on the battlefield (DoDIG, 2001; U.S. Army, 2011).

In this paper, we explored the potential benefits of a system under test (SUT), which improves the situational awareness on the battlefield by allowing soldiers to detect and identify threats (persons or otherwise) from a secure distance and in a reasonable timeframe. We comparatively examined the effect of the SUT on unit mission success, casualties, and fratricides relative to those units that do not possess the SUT based on data collected from a previous Limited User Test 2009 (LUT 09). Specifically, based on the means and standard deviations of the selected metrics collected in the LUT 09, we evaluated whether the behavior of these means and standard deviations can reasonably be expected to generate differences that are statistically significant during any subsequent Initial Operational Test & Evaluation (IOT&E) event.

Evaluation Metrics

A listing of the proposed evaluation paradigm for the IOT&E comparison was developed and approved by testing offices within the Department of Defense. From this listing, we examined select metrics for which data are available from the earlier LUT 09. First, we computed the means (or other measures of central tendencies) and the standard deviations. Next, assuming those values represented the “treatment” situation (i.e., as the SUT was used in the LUT 09, the “treatment” situation is the unit performance with the SUT), we examined how different the performance would have to be in the baseline for there to be a statistical difference exhibited in the data we have. In all cases, we assumed that the standard deviation exhibited in the baseline case is the same as that in the treatment situation as no variability data exists for the baseline. However, this assumption is tested later in our sensitivity analysis in the Sensitivity Analysis section. (*Note: Also, as the unit claimed that the systems did not help them, we reexamined the data assuming the values obtained in the test represent the “baseline” and observed how much improvement the new systems must provide in order to be different. In most cases, the magnitude of the difference is all that matters; thus, whether the data we have is baseline or treatment is a moot point except in select cases where the parameters are not symmetric*). The measures we examined were as follows:

- The number of times BLUFOR accomplished the assigned mission. (Note: We compared the *number* of battles, using a "sign test" for paired battles, and also the *percentage* of total battles that the BLUFOR accomplished.)
- The number of BLUFOR and OPFOR casualties. (Specifically, we looked at percentages.)
- The number of fratricides incidents. (Specifically, we looked at percentages.)

These three measures assess the *top level performance* of a unit relative to another unit. It is understood that the actual comparison during any subsequent IOT&E will look at numerous other measures, including subjective ones. For example, structured interviews may be considered complimentary to the statistical analysis and may be performed during the comparative test to aid in explaining why statistical differences did or did not occur during the test. However, for many of these measures, no data were collected in LUT 09, and we felt that the selected measures would give a reasonable indication of what to expect.

The key discriminators between the SUT battalion and the baseline battalion would be (1) the degree of improved situational awareness provided to the unit and (2) the impact of having this improved situational awareness. The expectation is that the metrics (or measures of merit) will show improved situational awareness attributable to the presence of the SUT, and no loss of lethality or force protection when compared to the baseline. The metrics are applicable to both the SUT and the baseline battalion and serve as the basis for comparison between them.

Definition of Metrics

Mission Success

Mission success is a complex measure driven by a number of factors, among which are the number of BLUFOR kills, the number of civilian kills, whether the unit achieved its objective, etc. For our analysis, we relied on expert determination by subject-matter experts (SMEs) on site during the test for identifying whether a mission was accomplished. We used two metrics to assess mission success. The first metric was the number of times the BLUFOR unit accomplished its missions. Using this metric, we performed a sign test to compare the baseline unit and the SUT unit. The second metric was the mission success rate. This metric normalizes the number of accomplished missions by the number of missions conducted. It is a bit more informative than simply the number of accomplished missions as it indicates the past success rate of the BLUFOR unit in accomplishing its missions. The mission success rate is calculated as follows:

$$Mission\ Success\ Rate = \frac{Number\ of\ Missions\ Accomplished}{Number\ of\ Missions\ Conducted} \quad (1)$$

For this metric, we used a two proportion *z*-test in the comparative analysis.

Casualties

In this study, we defined casualties as the number of kills a unit sustains. Initially, we considered two metrics to assess casualties sustained by the units. These were the number of losses and the casualty rate. The number of losses is the total number of casualties a unit incurs over the mission. While this metric gives a first order glimpse of the force protection capability of the unit, it does not account for the cost of these casualties to the unit. For example, a casualty loss of 20 soldiers is more costly to a unit that has a starting strength of 30 than it is to a unit that has a starting strength of 130. For this reason, we decided to look at the casualty rate. The casualty rate incorporates information about the cost of casualties

to the unit by normalizing the number of casualties a unit sustains by the unit starting strength. The formula for the casualty rate is shown below.

$$Unit\ Casualty\ Rate = \frac{Number\ of\ Losses\ Sustained\ by\ Unit}{Starting\ Strength\ of\ Unit} \quad (2)$$

Using the previous example, a unit with a starting strength of 30 which has 20 casualties will have a high casualty rate of 0.66, while a unit with a starting strength of 130 will have a low casualty rate of 0.15. In this analysis we used the casualty rate and the student *t*-test to draw conclusions on what to expect in the IOT&E.

Fratricides

A number of definitions exist for fratricides. Among these are (1) any engagement in which a friend fires at a friend, whether damage is done or not and 2) casualties caused by friendly fire. For the purpose of this analysis we used the second definition, casualties caused by friendly fire, as this definition more accurately reflects the damage caused. It is important to note, however, that the first definition is equally as relevant as the second definition in assessing how well the soldier is able to distinguish a threat from a friendly. In a similar vein to casualties, we considered two metrics, the number of fratricides and the fratricide rate. This rate is the number of unit fratricides as a percentage of the total unit casualties.

$$Unit\ Fratricide\ Rate = \frac{Number\ of\ Unit\ Fratricides}{Number\ of\ Unit\ Casualties} \quad (3)$$

For similar reasons discussed previously, we selected the fratricide rate as the comparison metric. This metric is insightful as it indicates the likelihood that a solider is killed by another soldier in the same unit. Alternatively, this metric may be viewed as a measure of the self-inflicted casualties in a unit. Using a student *t*-test, we sought to determine whether it is possible to draw statistical conclusions about the ability of the SUT to reduce fratricides relative to the baseline systems through improved situational awareness.

Potential Outcomes and Analysis

Mission Success

There were 13 missions in the SUT LUT 09: Three were attack missions, two were defend missions, three were cordon and search missions, four were raid missions, and the remaining one was a stability ops mission. The BLUFOR had an 85% success rate, accomplishing 11 of the 13 missions, partially accomplishing one, and failing to accomplish one. A summary of these mission success statistics is shown in Table 1.

Table 1. Description of Mission Outcomes

Mission	Mission Type	Successful	BLUFOR Starting Strength	BLUFOR Casualties	OPFOR Starting Strength	OPFOR Casualties
1	Raid	yes	130	10	50	26
2	Raid	yes	130	7	50	25
3	Defend	yes	130	25	50	0

4	Attack	yes	130	15	50	10
5	Attack	yes	130	25	50	8
6	Cordon and Search	yes	130	8	50	7
7	Defend	yes	130	16	50	15
8	Cordon and Search	yes	130	12	50	6
9	Raid	partially	130	7	50	3
10	Cordon and Search	yes	130	20	50	8
11	Attack	no	130	14	50	10
12	Stability Operations	yes	130	2	50	5
13	Raid	yes	130	10	50	22

An initial glance at the high mission success rate might imply that the SUT contributed positively to situational awareness and thus operational performance. However, caution is advised against prematurely drawing this conclusion from the data. Table 1 indicates that on average the BLUFOR starting strength was about two to three times that of the OPFOR. This difference in starting strength may give the BLUFOR a significant advantage. Without properly accounting for this advantage, it is possible to incorrectly conclude that the performance of the BLUFOR is attributed to the SUT.

Missions Accomplished

To evaluate the possibility of determining whether the high mission success rate may be attributed to SUT, we conducted statistical analyses on the LUT 09 data. In particular, we examined how many missions the baseline unit would have to lose, given the 85% mission success rate (or 11 missions accomplished) of SUT unit, to be significantly different statistically. The first metric examined was the number of missions accomplished. For the analysis of missions accomplished, we use the binomial sign test (Sheskin, 2004).

The binomial sign test compares differences in the performance of baseline systems relative to the SUT system using paired tests. The test statistic only considers the mission outcomes, which differ between the system under test and the baseline systems as these differing outcomes act as discriminators between the two tests. These possible outcomes are shown in Table 2. Only 12 missions were evaluated, as the partially successful mission was eliminated from the dataset. The variables in the table were defined as follows:

$\mathbf{X_{WW}}$ - number of missions accomplished by both the baseline and the SUT units

$\mathbf{X_{WL}}$ - number of missions accomplished by the SUT unit, but not the baseline unit

$\mathbf{X_{LW}}$ - number of missions accomplished by the baseline unit, but not the SUT unit

$\mathbf{X_{LL}}$ - the number of missions not accomplished by both the baseline and the SUT unit

Table 2. Notional Representation of Mission Outcomes

	Baseline	
	W	**L**

SUT	**W**	X_{WW}	X_{WL}	11
	L	X_{LW}	X_{LL}	1
		$X_{WW} + X_{LW}$	$X_{WL} + X_{LL}$	12

For the binomial sign test, the number of trials was defined as $X_{LW} + X_{WL}$, or the total number of missions in which the outcome differs between the two groups. The number of successes was defined as X_{WL}, or the number of times the SUT unit outperforms the baseline unit. The null hypothesis assumed that there is no difference between the mission performance of the SUT unit and the baseline unit. Therefore, a success and a non-success are equally likely to occur, leading to the null hypothesis being defined as:

$$H_{\varnothing}: p = 0.5 \quad (4)$$

where *p* is the likelihood of success. As the objective of the comparative test during any subsequent IOT&E is to determine whether the SUT positively contributes to situational awareness, our alternative hypothesis was one directional and given by:

$$H_a: p > 0.5 \quad (5)$$

The test was performed with a 90% confidence level. Next, we used the cumulative binomial probability distribution function to determine the likelihood that the SUT unit outperforms the baseline unit a certain number of times or greater (e.g., eight or more times). If this probability was less than $\alpha = 0.1$, we rejected the null hypothesis and concluded that there is statistical evidence the SUT unit outperforms the baseline unit and enhances situational awareness. For this study, we were concerned with the number of losses by the baseline unit for statistical significance. Table 3 shows the results of the analysis where the probability columns indicate the likelihood that the SUT unit outperforms the baseline unit by at least a certain number of missions (e.g., eight or more), and the required baseline losses columns indicate the actual losses required to observe the SUT unit outperform the baseline unit by a certain number of missions.

Table 3. Required Baseline Losses

		X_{LW} = 1		X_{LW} = 0	
		Pr(Number of Successes $\geq X_{WL}$)	Required Baseline Losses	Pr(Number of Successes $\geq X_{WL}$)	Required Baseline Losses
X_{WL}	0	1.000	0	1.000	1
	1	0.750	1	0.500	2
	2	0.500	2	0.250	3
	3	0.313	3	0.125	4
	4	0.188	4	0.063	5
	5	0.109	5	0.031	6
	6	0.062	6	0.016	7
	7	0.035	7	0.008	8
	8	0.020	8	0.004	9
	9	0.011	9	0.002	10
	10	0.006	10	0.001	11
	11	0.003	11	0.000	12

There are a couple of facts to note about this table. First, as the SUT unit failed to accomplished only one mission, there are only two possible values for X_{LW}. This greatly reduced our analysis. However, a single loss meant that there were 12 possible values for X_{WL}, all of which are laid out in the table. As expected, the probability of achieving a certain number of successes or greater decreased as X_{WL} increased. For example, assuming the baseline unit outperforms the SUT unit in one mission (i.e., $X_{LW} = 1$), the probability of the SUT unit outperforming the baseline unit six or more times is 0.062, while 11 or more times is 0.003. The table also indicates that the required minimum number of losses by the baseline unit for statistical significance is six, or around half of the 12 missions. Five losses or lower will lead to statistically inconclusive results. In the case where the baseline never outperforms the SUT unit (i.e., $X_{LW} = 0$), the minimal number of losses to show a statistically significant difference between the two units is five. Four or more losses will lead to statistically inconclusive results.

In summary, based on the LUT 09 results in which the BLUFOR won 11 of the 13 battles, a baseline unit would have to underperform the SUT unit by five or more missions to be considered statistically different from the observed outcome of the SUT unit.

Mission Success Rate

The second metric of mission success considered was the mission success rate. We used a one-tail two-proportion *z*-test to determine the required reduced level of baseline unit performance to produce a statistically significant difference. The following formula was applied (Ott & Longnecker, 2010):

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \quad (6)$$

where $\hat{p}_1$ is the success rate of the SUT unit and $\hat{p}_2$ is the success rate of the baseline unit and n_1 and n_2 is the number of missions conducted by the SUT and the baseline unit, respectively. Using a *z*-value of 1.28 and a mean SUT mission success rate of 0.85 (or 85%), we solved for the baseline mean mission success rate given the number of baseline missions conducted. Figure 3 shows the missions success rate and the number of missions conducted. The mean mission success rate depicted in the figure is the maximum rate the baseline unit can achieve. Beyond this maximum rate, the results of the comparative evaluation become statistically inconclusive. The red dot is the mission success rate of the SUT unit observed in LUT 09.

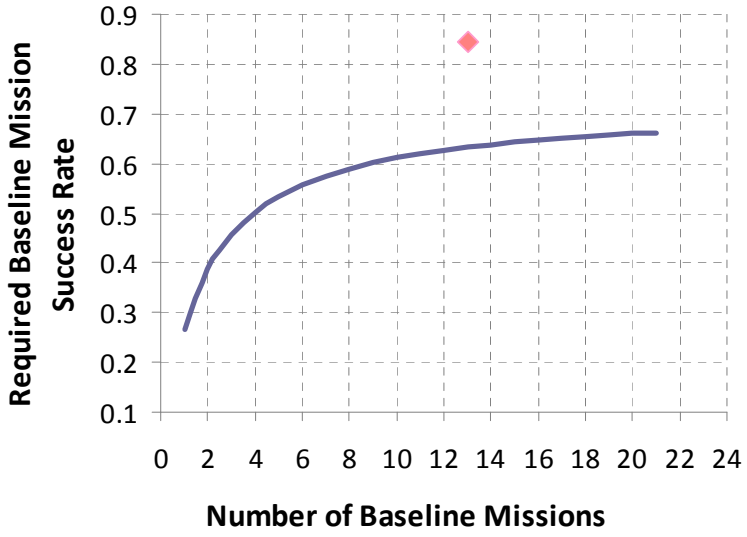


Figure 3. Maximum Required Mean Baseline Mission Success Rate

The figure shows that for a large number of baseline missions conducted, for example 20 missions, the required mean mission success rate for statistical difference is approximately 66%. As the number of missions decreases to 13 or that were conducted in the LUT 09, the required mission success rate is 63%. Further decreases in the number of missions conducted result in success rates below 60% (i.e., there is not a great deal of confidence that the baseline unit will accomplish the mission).

Figure 4 breaks down the mission success rate by showing the number of the required mission losses given the number of missions conducted. If 13 missions are conducted, the required number of baseline losses for statistical significance is five. It is interesting to note, that this closely mirrors the results obtained from using the binomial sign test. That is, whether considering the sign test (number of missions accomplished) or mission success rate, in order for there to be a significant difference between a baseline and SUT unit, the baseline unit needs to lose four to five more missions than the SUT unit based on the LUT 09 data.

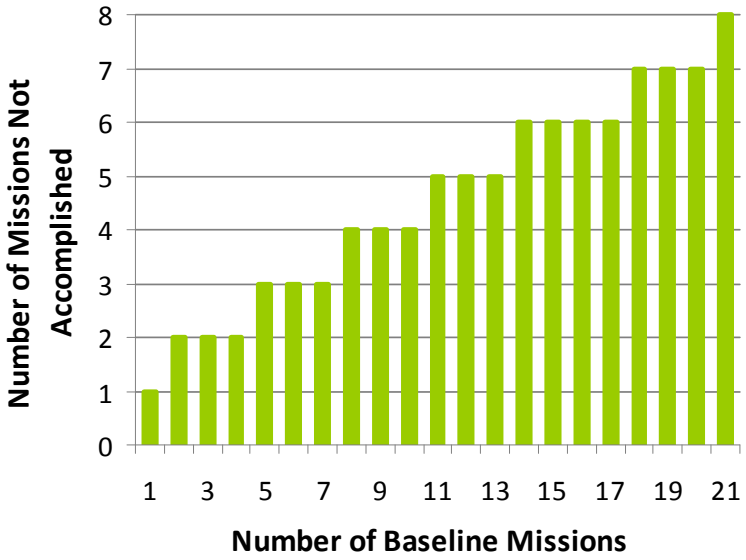


Figure 1. Minimum Number of Baseline Mission Losses

For both metrics of mission success considered, the statistics indicated that the baseline unit will need to perform very poorly to produce conclusive results. However, as currently constructed, the overwhelming starting strength of the BLUFOR argues against such an outcome. The BLUFOR starting strength to the OPFOR starting strength ratio is on average greater than 2:1. This provides the BLUFOR with a significant advantage, which may be leveraged to overwhelm the OPFOR during various operations with or without the SUT. Given the current test set-up of a substantial BLUFOR manpower advantage relative to the OPFOR manpower, it is unlikely that these metrics will produce any conclusive results about the contribution of the SUT to mission success by enhancing situational awareness and understanding in a comparative evaluation.

Mission Casualties

One of the metrics used to judge whether the SUT provides better situational awareness was a decline in BLUFOR casualties during tactical battles. Figure 5 shows casualty data per mission for LUT 09.

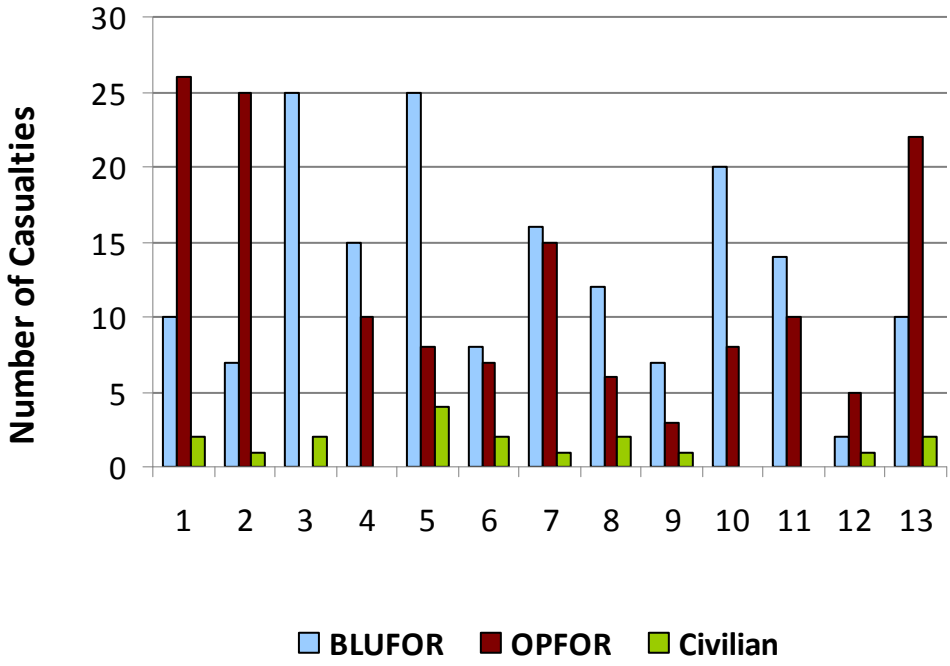


Figure 5. Number of Casualties per Mission

There was a high number of BLUFOR casualties in almost every operation, with the greatest number of casualties occurring primarily in attack and defend missions. For the OPFOR, high losses were incurred primarily during raids with Mission 1 being the most devastating. Although not shown on this figure, it is interesting to note that despite the potential increase in situational awareness offered by the SUT system, all civilian casualties were caused by the BLUFOR.

In order to assess whether there was a plausible likelihood of drawing statistical conclusions about the difference in casualties sustained by the SUT unit and those sustained by the baseline unit, we used the casualty rate. This metric was defined previously as the following:

$$Unit\ Casualty\ Rate = \frac{Number\ of\ \ Losses\ Sustained\ by\ Unit}{Starting\ \ Strength\ of\ \ Unit} \tag{7}$$

Next we used a one-directional student *t*-test at the 90% confidence level to determine the minimum required mean casualty rate of the baseline unit to generate a statistical difference. The formula we used for the student *t*-test statistic is shown below (Sheskin, 2004):

$$t_{\alpha,df} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{8}$$

where $\overline{X}_1$ is the mean casualty rate of the SUT unit and $\overline{X}_2$ is the mean casualty rate of the baseline unit, n_1+n_2-2 is the degrees of freedom and S_1 and S_2 are the standard deviations for the respective units. We solved for the baseline casualty rate given

an average SUT unit casualty rate of 10.1% and the number of baseline missions conducted. Figure 6 displays the results. The mission success rate depicted in the figure is the minimum casualty rate the baseline unit can achieve for significance. Lower than this rate, the results of the comparative evaluation become inconclusive. The red dot is the casualty rate of the SUT unit.

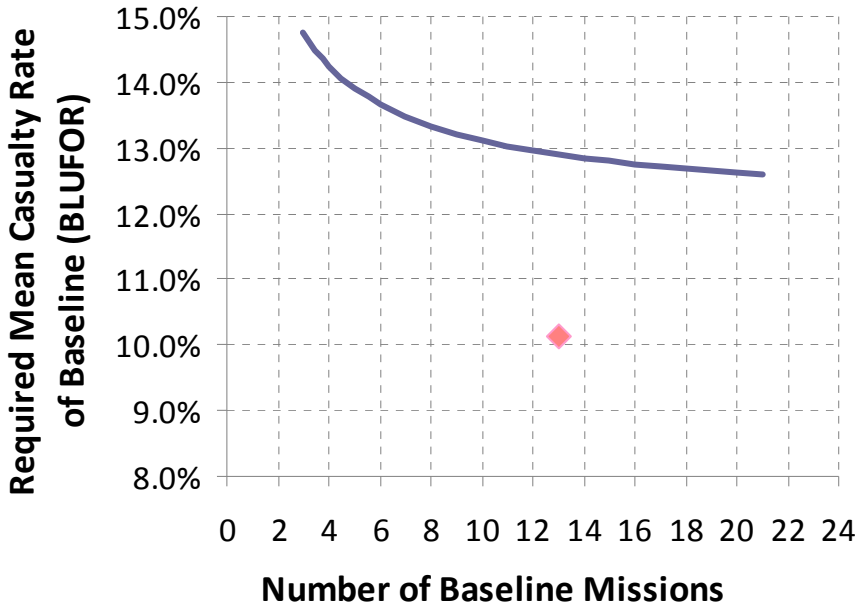


Figure 6. Minimum Required Mean BLUFOR Casualty Rate

The figure shows that for a large number of baseline missions conducted, for example 20 missions, the minimum required casualty rate for statistical difference is 12.6%. As the number of missions decreases to 13 or that were conducted in the LUT 09, the minimum required casualty rate is 12.9%. Further decreases in the number of missions conducted results in higher required casualty rates for the baseline BLUFOR.

We conducted a similar analysis for the OPFOR, the results of which are shown in Figure 7. For the OPFOR, the two units being compared are the OPFOR unit against an SUT unit, and the OPFOR unit against a baseline unit. In contrast to the BLUFOR, the mean casualty rate shown in Figure 7 is the maximum casualty rate sustained by the OPFOR for a given number of missions. The red dot is the mean casualty rate of the OPFOR unit against the SUT unit and has a value of 22.3%.

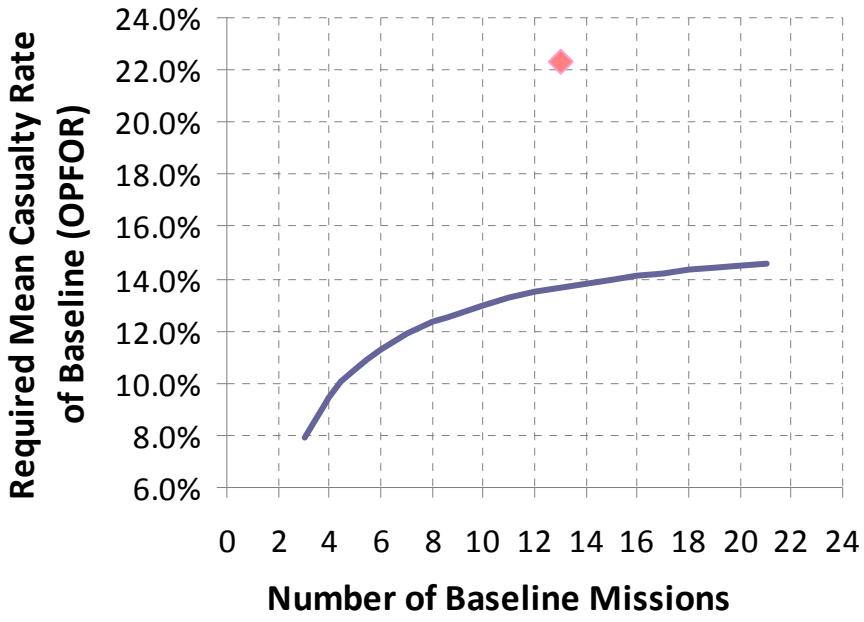


Figure 7. Maximum Required Mean OPFOR Casualty Rate

The results indicate that a maximum casualty rate of 14.5% is required if the IOT&E conducts 20 or more baseline missions. Conducting approximately 13 baseline missions will necessitate a maximum casualty rate of 13.7% for a statistical difference. As the number of missions decreases below 13, the maximum required casualty rate falls to low values of 12.0% or below. That is, for a starting strength of about 60, the OPFOR unit will only lose about seven soldiers.

In order to determine whether the required mean OPFOR and BLUFOR casualty rates are reasonable values, we compared these values to rates generally observed from tactical battles in operational tests. We selected an average casualty rate of approximately 10% as our guideline based on discussions with analysts from the Institute of Defense Analyses. The value of 10% was in the range of that exhibited by the SUT unit during the LUT 09. Using the 10% guideline, we surmised that achieving a required mean BLUFOR casualty rate of 12.9% may be possible during IOT&E. In the case of the OPFOR, traditional casualty rates have been on the order of 25% in recent operational tests. Therefore, achieving a casualty rate of below 16% may prove quite challenging during IOT&E. While there is a possibility of obtaining significant results for the BLUFOR casualty data, the low OPFOR casualty rate required makes it unlikely that the current test set-up will yield statistically conclusive results in the case of the OPFOR. In other words, it is possible to observe the SUT contribute to a reduction in BLUFOR casualty during IOT&E but highly unlikely to observe a contribution to BLUFOR lethality.

Mission Fratricides

One of the expected effects of better situational awareness is reduced BLUFOR fratricide. Figure 8 shows the number of fratricide incidents per mission for the LUT 09.

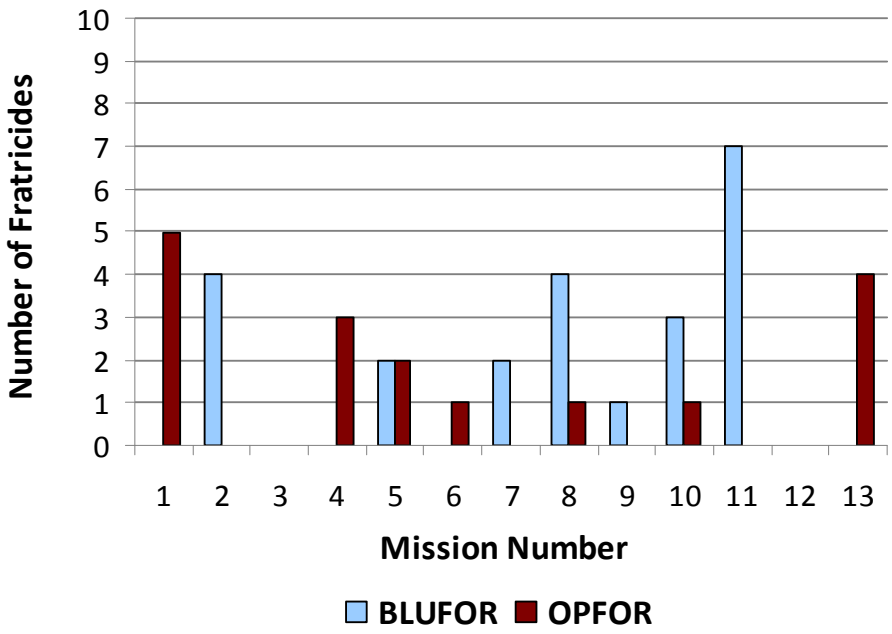


Figure 8. Number of Fratricide Incidents per Mission

In Missions 2, 8, 10, and 11, the BLUFOR sustained a large number of fratricides relative to the other missions. These missions were diverse in type with Mission 2 being a raid, Missions 8 and 10 being cordon and search missions, and Mission 11 being an attack mission. As such, no initial conclusions may be drawn about the tendency of certain missions to produce BLUFOR fratricides. The OPFOR sustained high fratricide losses in three missions, two of which were raid missions.

In this analysis, we were concerned primarily with BLUFOR fratricides. As OPFOR does not possess the SUT, OPFOR will not have enhanced situational awareness. It is expected the SUT will have no effect on the OPFOR fratricides. We used the fratricide rate to determine the limits for statistical significance. As stated previously, the fratricide rate is defined as the following:

$$Unit\ Fratricide\ Rate = \frac{Number\ of\ Unit\ Fratricides}{Number\ of\ Unit\ Casualties} \tag{9}$$

Similar to the casualty rate, we implemented a one-directional student t-test at the 90% confidence level to determine the minimum required fratricide rate of the baseline unit to generate a statistical difference. We solved for the baseline fratricide rate given an average SUT fratricide rate of 14.6% and the number of baseline missions conducted. Figure 9 displays the results. The fratricide rate depicted in the figure is the permissible minimum rate the baseline unit can achieve. Beyond this rate, the results of the comparative evaluation become inconclusive. The red dot is the mean fratricide rate of the SUT unit.

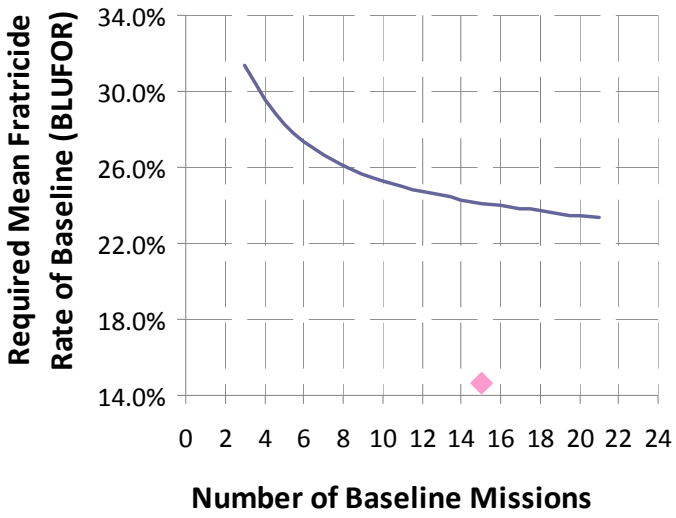


Figure 9. Minimum Required BLUFOR Fratricide Rate

The analysis suggests that for an extremely low number of missions completed (e.g., eight or below), the minimum required fratricide rate surpasses 26.0%. This high rate implies that approximately a quarter of all casualties sustained by the BLUFOR would need to be self inflicted. As the number of missions conducted approaches 13 or that were conducted in the LUT 09, the required fratricide rate falls below 26.0% to approximately 25.0%. Conducting additional missions will only have a marginal effect on the minimum required fratricide rate as, at 21 missions conducted, the rate falls to 23.4%.

In order to gain an idea of whether these required fratricide rates are reasonable, we first needed some idea of the average fratricide rate in actual tactical operations. Precise fratricide data is relatively difficult to obtain. However, available reports placed the fratricide rates around 13% (Bower, Lacey, & McCarthy, 2003; Gadsden & Outteridge, 2006). Using this figure as a guideline we determined the plausibility of drawing any statistical conclusions from the comparative tests. The minimum mean required fratricide rate shown in Figure 9 is almost 100% higher than that experienced during actual tactical operational conditions. Given this high required BLUFOR fratricide rate, it is unlikely that the current test set-up will yield statistically conclusive results.

One additional point is worth noting regarding fratricides. First, the exponential-like nature of the slope in Figure 9 suggests there are diminishing returns to conducting an increasing number of missions. If we extended the analysis to 100 conducted missions, the minimum required fratricide rate will remain relatively high at 21.0%. Thus, the statistical returns from conducting a large number of missions may not justify the incremental test costs.

Sensitivity Analysis

The analysis performed in the previous sections is predicated on a number of assumptions. Among these assumptions are (1) the variability in performance measures across missions is identical for both the SUT unit and the baseline unit, (2) 90% confidence level is the more appropriate confidence level for the statistical analysis, and (3) the performance of the SUT unit in the LUT 09 is representative of its future performance in subsequent IOT&E. For the sensitivity analysis, we modified each of these assumptions and examined the corresponding response of the required unit mean performance limits for statistical significance.

One key assumption in the previous analysis was that the variability exhibited in several of the SUT performance measures may be used as a proxy for the variability of the baseline unit performance. We believed this assumption to be justifiable as this has been the case in many prior side-by-side tests. In order to test the robustness of our conclusions to changes in variability, we relaxed our assumption by assuming the performance variability of the baseline unit is half that of the SUT unit.

Confidence levels are often used to establish bounds on performance metrics in the presence of uncertainty. While traditionally analysis is conducted at standard confidence levels (90%, 95%, or 99%), the criteria for selecting confidence levels are arbitrary. To understand the sensitivity of our previous analysis to the changes in confidence levels, we reduced the confidence interval to 80%.

Finally, survey results in the LUT 09 indicated that the unit claimed the SUT was not instrumental in accomplishing missions. Based on this response, we reexamined the data assuming the values obtained in the LUT 09 represent the baseline unit instead of the SUT unit. Next, we evaluated how much improvement the new system must provide in order to be statistically different.

Table 4 shows the results for the sensitivity analysis for the 13 conducted missions in LUT 09. The initial results are those obtained in the previous sections (i.e., initial results assumed the LUT 09 was representative of the SUT and described how well or poorly a baseline must perform to be different from the SUT).

Table 4. Results of Sensitivity Analysis

		Required Values for Statistical Significance in IOT&E			
Metrics	**Observed in LUT 09**	**Initial Results**	**50% Variability Reduction**	**Confidence Level = 80%**	**SUT**
Missions Not Accomplished	1	4-6	N/A	4	--
Mission Success Rate	0.85	63.2%	N/A	71.1%	98.2%
BLUFOR Casualty Rate	10.1%	12.9%	12.1%	11.9%	4.7%
OPFOR Casualty Rate	22.3%	13.7%	16.2%	16.7%	31.0%
BLUFOR Fratricide Rate	14.6%	24.5%	21.6%	21.2%	7.3%

For the mission success metrics (missions losses and success rate), only two of the three scenarios applied. Relaxing the confidence level to 80% saw the number of mission losses by the baseline unit fall by about one or two to four (for both cases in which $X_{LW} = 0$ or $X_{LW} = 1$) or around 30.8% of all missions conducted. The mission success rate of the baseline unit increased to 71.1% from 63.2%. Interestingly, reducing the variability by 50% or the confidence level to 80% exhibited almost identical impacts on the casualty and fratricide rates. In both cases, the BLUFOR casualty rate declined to just over 11.5%, the OPFOR casualty rate rose to just over 16.0% and the BLUFOR fratricide rate fell to around 21.0%.

There are a number of insights that we can draw from the sensitivity analysis. First, it is unlikely that the comparative test at the IOT&E will lead to any statistically relevant conclusion regarding the BLUFOR fratricide rate. Under relaxed assumptions, the fratricide rate needed to show a statistical difference remained high, above what is currently observed in combat (Bower, Lacey, & McCarthy, 2003; Gadsden & Outteridge, 2006). Second, relaxing the variability by half and shifting the confidence interval from 90% to 80% provide

results that are more consistent with the hoped for operational performance of the baseline unit in the cases of the mission success and BLUFOR casualty measures. The OPFOR casualty measure remained significantly lower than that normally observed in actual combat. Judging from the new mission and BLUFOR casualty rates, one might infer that there may be some opportunity to produce statistically conclusive results for the mission success and BLUFOR casualty metrics

The sensitivity analysis provided a potentially positive outlook for gaining conclusive information about the ability of the SUT to enhance situational awareness evident through reduced casualties and improved mission success. The potential outlook may support the argument for conducting a comparative test as planned. However, it is important to understand the disadvantage of relaxing these two assumptions. Operational tests are often complex with a great degree of variability in test parameters. While relaxing the assumptions produces plausible results, there is a lower degree of confidence associated with the derived conclusions. Rephrased in a more colloquial manner, increasing the likelihood of drawing conclusions reduces the confidence in those conclusions.

Recall in the third set of sensitivity analyses, we assumed that data gathered from the LUT 09 was representative of the baseline unit as opposed to the SUT unit. This was done as units noted that the SUT did not help in their missions. From the results of the sensitivity analysis, and if we assumed that the performance exhibited in LUT 09 formed a baseline, the mission success rate indicates that the SUT unit would effectively need to win all of their missions in order to produce statistically significant conclusions. However, a review of the metric and the number of mission losses by the SUT unit suggests that it is not possible to obtain statistically significant results even with a 100% mission success rate. The minimum required mean SUT BLUFOR casualty rate was extremely low at 4.7% and the maximum required SUT fratricide rate to yield statistical significance at 7.3%. These rates imply for a unit with a starting strength of 140, only approximately seven casualties and zero fratricides occur.

The conflicting inferences drawn about the SUT unit needed improvement in regard to the mission success measures raised questions about whether it is possible to draw statistical conclusions given the current set-up of the side-by-side test. The required mean casualty and fratricide rates appear to be optimistic. At this point, we reserved any judgment about the possibility of the SUT unit achieving such rates. It is possible that there will be significant performance improvements in the SUT system during subsequent testing.

Conclusion

The objective of this analysis was to demonstrate the use of statistical inference to better understand the potential of an SUT in improving the operational performance of a given unit prior to conducting a comparative test. Specifically, we wanted to establish expectations for the statistical outcome of a comparative test by examining whether the behavior of the means and standard deviations of the metrics can reasonably be expected to generate differences. Initial results indicated that while the comparative test set-up for the SUT may yield statistically significant results for one of the system evaluation metrics, it is highly unlikely that evaluators will observe statistical significance in the remaining metrics.

One factor often cited for the lack of observed statistical significance is large variability in performance metrics due to the complexity of operational tests. While there is some justification for this statement, there are other factors, which may be adjusted in the test structuring to yield a more informative test outcome. By dissecting the underlying factors, which drive each metric, we pointed to potential improvements for test structuring, which may enhance the likelihood of observing differences in a greater percentage of the

metrics. Most notably was a recommendation to reconsider the BLUFOR to OPFOR starting strength ratio. While this ratio may be representative of current field operations, it is somewhat inhibitive to understanding the potential benefits of the SUT to unit performance.

Finally, the metrics presented in this analysis provide a glimpse into the top level performance of a unit with the SUT. However, it is important to note that the analysis did not consider qualitative measures of operational effectiveness. Information for qualitative assessments is gathered through surveys and structured interviews, and it may provide added insights not immediately evident in the quantitative metrics used.

References

Bower, A., Lacey, J., & McCarthy, T. (2003, April 7). Fratricide: Misfiring in the fog. *Time Magazine.*

Commission on Behavioral and Social Sciences and Education (CBASSE). (1995). *Statistical methods for testing and evaluating defense systems: Interim report*. Washington, DC: National Academies Press.

Commission on Behavioral and Social Sciences and Education (CBASSE). (1998). *Statistics, testing, and defense acquisition: New approaches and methodological improvements.* Washington, DC: National Academies Press.

Department of Defense Inspector General (DoDIG). (2001, March 30). *Acquisition of the Battlefield Combat Identification System* (Report No. D-2001-093).Washington, DC: Author.

Fox, B., Boito, M., Graser, J. C., & Younossi, O. (2004). *Test and evaluation trends and costs for aircraft and guided weapons* (Report No. MG-109). Santa Monica, CA: RAND.

Gadsden, J., & Outteridge, C. (2006, August 29–September 1). What value analysis? The historical record of fractricide. *23rd International Symposium on Military Operational Research.*

Ott, R. L., & Longnecker, M. (2010). *An introduction to statistical methods and data analysis* (6th ed.). Florence, KY: Brooks/Cole Cengage Learning.

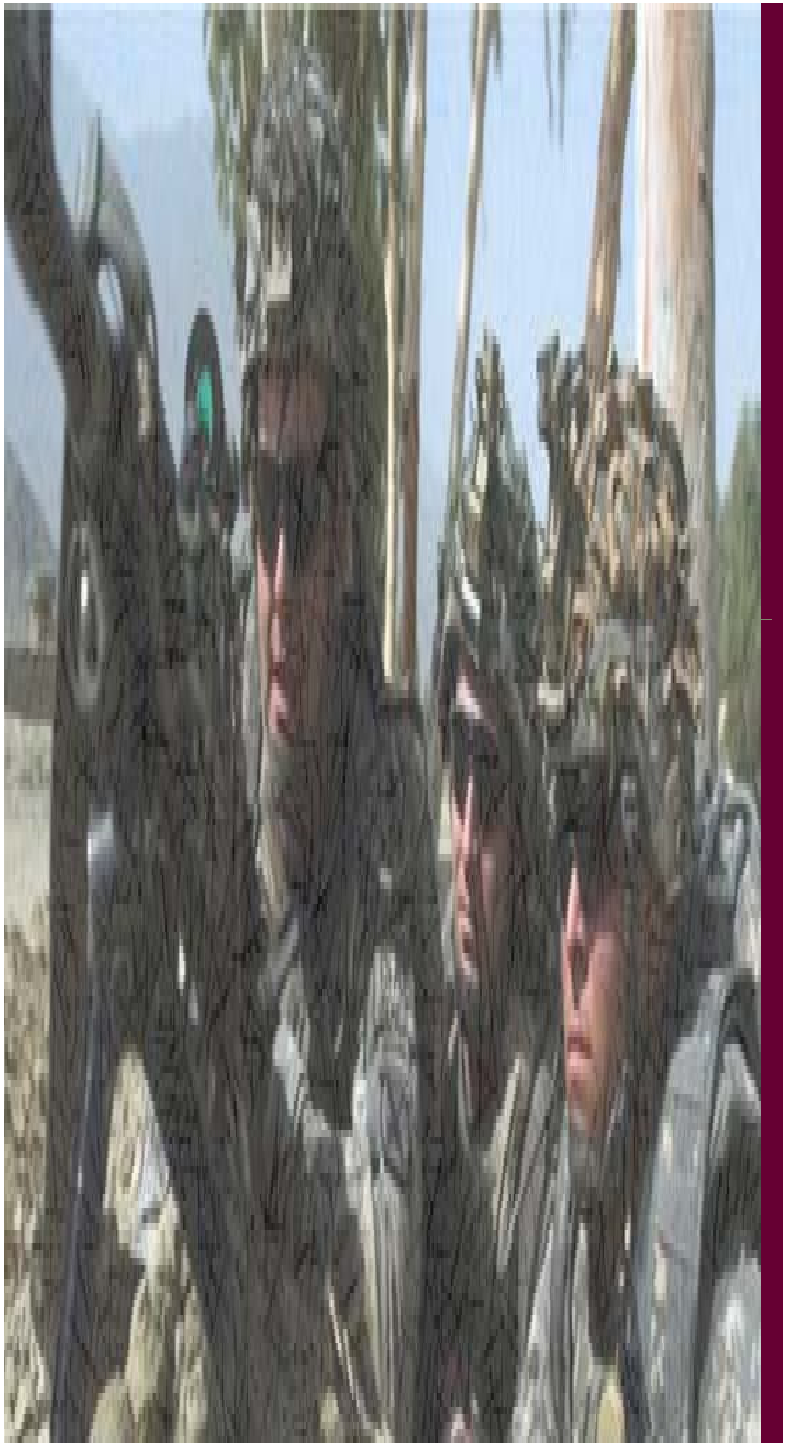
Schwartz, M. (2010, April 23). *Defense acquisitions: How DoD acquires weapon systems and recent efforts to reform the process*. Washington, DC: Congressional Research Service.

Sheskin, D. J. (2004). *Handbook of parametric and nonparametric statistical procedures* (3rd ed.). New York, NY: Chapman and Hall.

U.S. Army. (2011, January 25). Army modernization: Early Infantry Brigade Combat Team (E-IBCT). Retrieved from http://www.bctmod.army.mil/EIBCT/index.html

USD(AT&L). (2008, December 8). *Operation of the defense acquisition system* (DoDI 5000.02). Washington, DC: Author.

Wooldridge, J. M. (2003). *Introductory econometrics: A modern approach*. Florence, KY: Thompson Learning.



Utilizing Statistical Inference to Guide Expectations and Test Structuring during Operational Testing and Evaluation

Joy Brathwaite
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Alton Wallace
Dr. Robert Holcomb
Institute for Defense Analyses

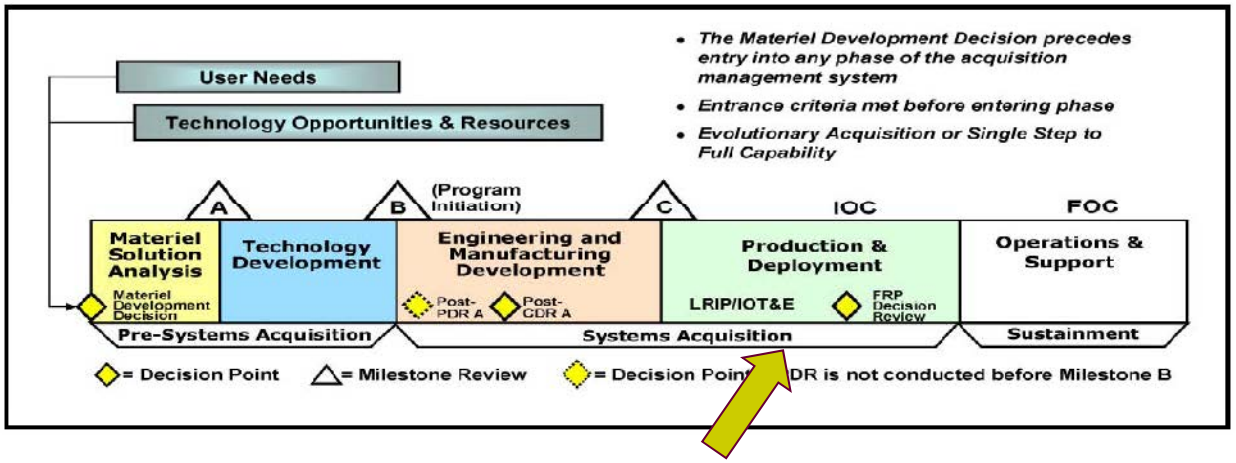
May 10-12, 2011

8th Annual Acquisition Research Symposium

Outline of Presentation

- Motivation
- Statistical Inference and Operational Testing
- Evaluating Potential Test Results
- Application to Situational Awareness System
- Summary

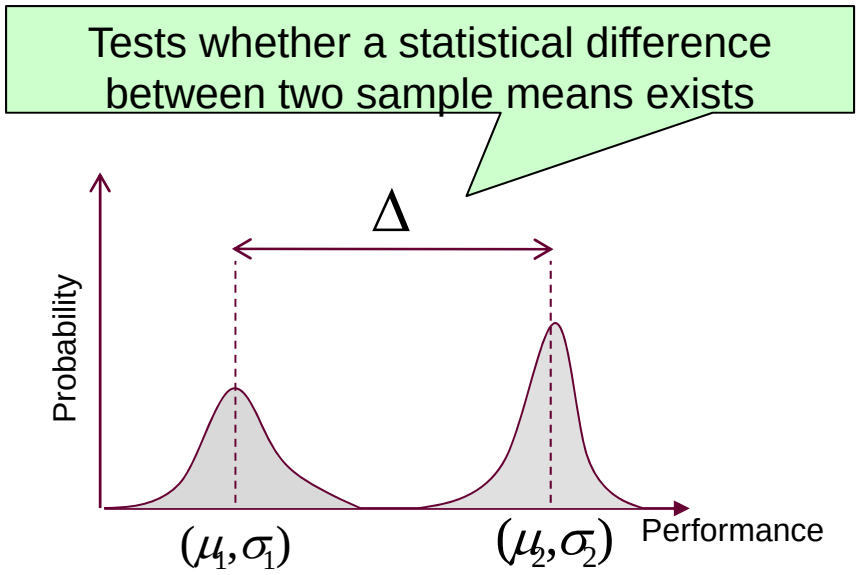
Motivation



- Initial Operational Testing and Evaluation occurs during the Production & Deployment acquisition phase
- Congress requires testing of major weapons systems to be conducted under operationally realistic conditions to determine operational suitability
- Comparative tests are utilized during operational testing to baseline a system under test (SUT) through a series of tactical battles
  - Goal is to determine whether and by how much the unit's performance systematically improves with the SUT
  - Several approaches, both quantitative and qualitative, are used to assess a systematic improvement (e.g. statistical analysis and user evaluations)

Statistical Inference

- Statistical inference noted as a best practice in system evaluation (CBASSE 1998)
- An applied statistical approach is often used to quantify and evaluate differences between treatment and control groups (Woolbridge 2003)
- In operational testing, statistical inference evaluates the performance difference between the SUT and the current status quo



Interval/Ratio Data	
Two independent samples	**t-test** z-test single factor between subjects ANOVA
Two dependent samples	t-test z test single factor between subjects ANOVA
Ordinal/Rank-Order Data	
Two independent samples	Mann-Whitney U test van der Waerden normal-scores test
Two dependent samples	Wilcoxon matched pairs signed-ranks test **Binomial sign test**
Categorical/Nominal Data	
Two independent samples	Chi-square test z-test
Two dependent samples	McNemar test Gart test

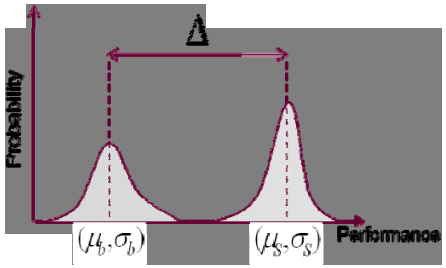
Statistical Inference

1 **State Research Question**

Does use of the SUT improve the mean performance of a unit?

2 **Specify Null and Alternative Hypotheses**

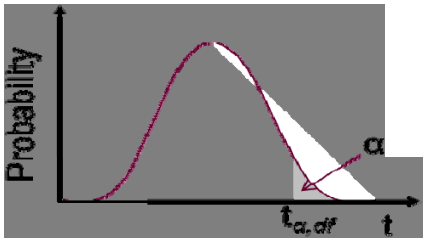
$$H_\phi : \mu_S = \mu_b$$
$$H_a : \mu_S > \mu_b$$



3 **Calculate Test Statistic**

$$t_{\alpha, df} = \frac{\overline{X}_S - \overline{X}_b}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_b^2}{n_b}}}$$

4 **Compute Probability of Rejection**



5 **State Conclusions**

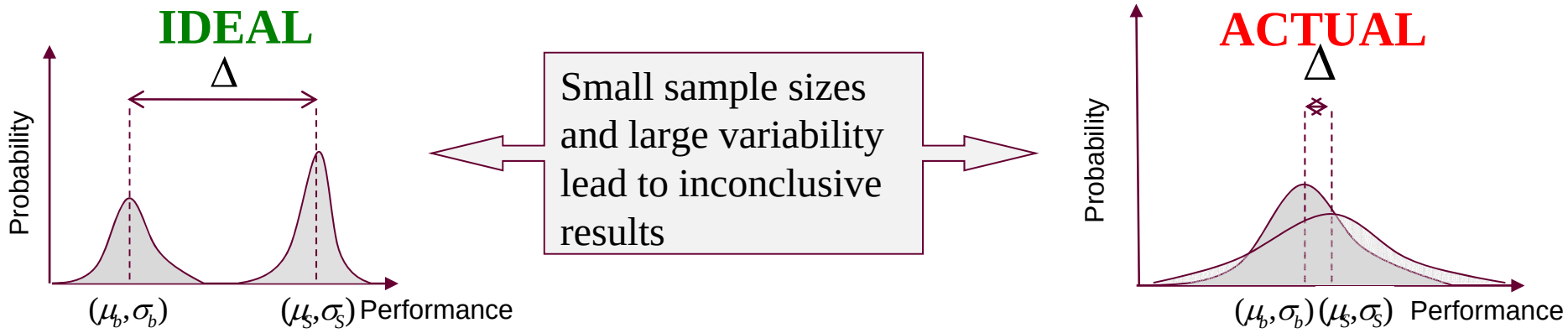
Did the SUT unit outperform the baseline unit statistically?

Statistical Inference in OT&E

- Evaluated a Situational Awareness System as an effective tool against fratricide in 2001 (Edwards 2001)
  - ➢**System Confidence Demonstration (SCD)**
    - No significant statistical difference between SUT and non-SUT units
    - Nearly impossible for SUT crew to statistically outperform baseline as baseline did so well
  - ➢**Virtual Integration Exercise (VIE)**
    - Overall, no significant difference occurred in fratricide rates between baseline and SUT

Assessing the difference in performance mean between two independent samples

$$t_{\alpha,\nu} = \frac{\overline{X}_S - \overline{X}_b}{\sqrt{\frac{s_S^2}{n_S} + \frac{s_b^2}{n_b}}}$$



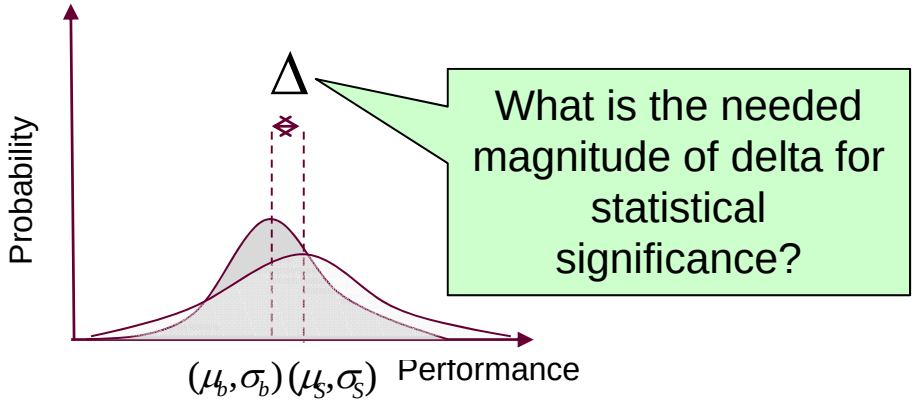
Evaluating Potential Test Results

- Comparative tests are costly to administer and difficult to repeat
- Understand potential results a priori to guide expectations, test structuring and enable a more effective utilization of resources

1. What improvement in the mean performance is needed over the baseline to confidently assess whether there is a statistical difference?
2. Is the required performance of the unit needed to show a statistical difference reasonable?

$$\overline{X}_b = \overline{X}_S - t_{\alpha,\nu}\left(\sqrt{\frac{s_S^2}{n_S} + \frac{s_b^2}{n_b}}\right)$$

ACTUAL



Guiding Expectations and Test Structuring

Analysis of Systematic Difference

- Several approaches, both quantitative and qualitative, are used to assess a systematic improvement (e.g. statistical analysis and user evaluations)
- Statistical inference noted as a best practice in system evaluation (CBASSE 1998)

Potential results of test a priori may:

- Provide guidance on the potential benefits of conducting test
- Provide guidance on structuring the test
- Lead to a more cost-effective test execution
- Provide maximal information given resources expended

Problems Experience in Previous Tests

- Evaluated a Situational Awareness System as an effective tool against fratricide in 2001 (Edwards 2001)
  - **System Confidence Demonstration (SCD)**
    - No significant statistical difference between SUT and non-SUT units
    - Nearly impossible for SUT crew to statistically outperform baseline as baseline did so well
  - **Virtual Integration Exercise (VIE)**
    - Overall, no significant difference occurred in fratricide rates between baseline and SUT

Outline of Presentation

- Motivation
- Statistical Inference and Operational Testing
- Evaluating Potential Test Results
- Application to Situational Awareness System
- Summary

Examination of the Force Effectiveness

- Operational needs statements from theater called for ground and aerial robotic capability to enable better situational awareness
- Evaluation of a SUT to improve the unit situational awareness on the battlefield
- Data on SUT performance gathered from its LUT 09
- Operational performance evaluation of a battalion with and without the SUT systems

Mission	Mission Type	Success	BLUFOR Starting Strength	BLUFOR Casualties	OPFOR Starting Strength	OPFOR Casualties
1	Raid	yes	130	10	50	26
2	Raid	yes	130	7	50	25
3	Defend	yes	130	25	50	0
4	Attack	yes	130	15	50	10
5	Attack	yes	130	25	50	8
6	Cordon and Search	yes	130	8	50	7
7	Defend	yes	130	16	50	15
8	Cordon and Search	yes	130	12	50	6
9	Raid	partially	130	7	50	3
10	Cordon and Search	yes	130	20	50	8
11	Attack	no	130	14	50	10
12	Stability Operations	yes	130	2	50	5
13	Raid	yes	130	10	50	22

Performance Metrics of Interest

1. **Missions Not Accomplished** → Considered missions which had a conclusive result

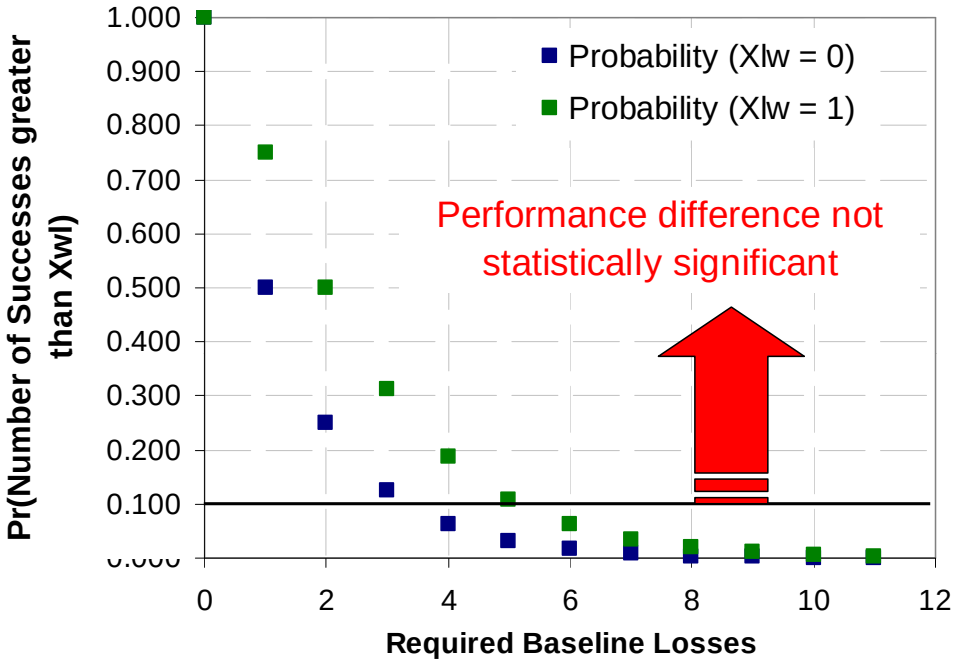
2. **Mission Success Rate** → $$MSR = \frac{Number\ of\ Missions\ Accomplished}{Total\ Missions\ Conducted}$$

3. **BLUFOR Casualty Rate** → $$B\,Casualty\ Rate = \frac{BLUFOR\ Losses}{BLUFOR\ Starting\ Strength}$$

4. **OPFOR Casualty Rate** → $$O\,Casualty\ Rate = \frac{OPFOR\ Losses}{OPFOR\ Starting\ Strength}$$

5. **BLUFOR Fratricide Rate** → $$Fratricide\ Rate = \frac{BLUFOR\ Fratricides}{BLUFOR\ Losses}$$

Missions Not Accomplished



Xwl – Number of missions accomplished by SUT unit but not baseline unit

Xlw – Number of missions accomplished by baseline unit but not SUT unit

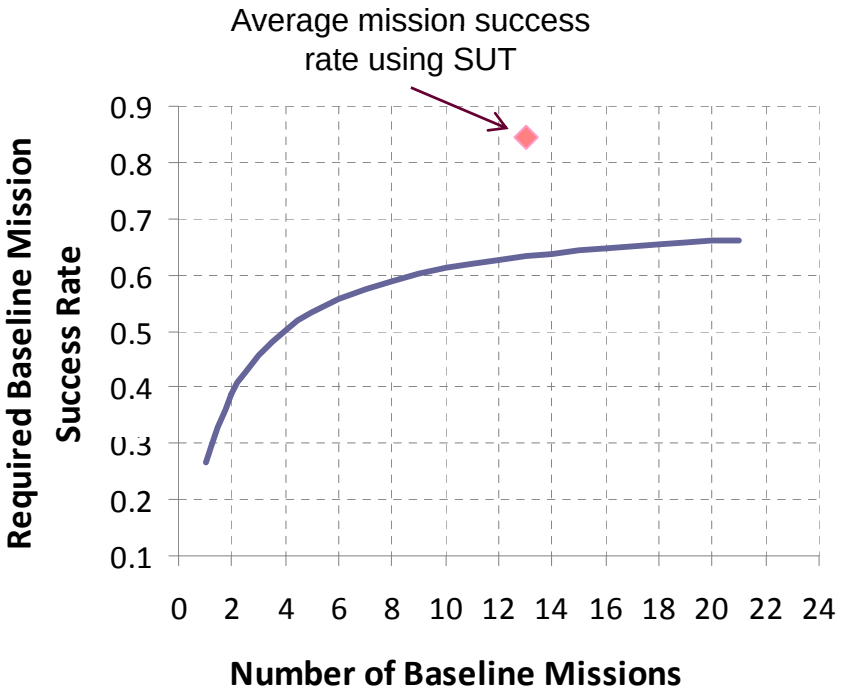
- Comparative evaluation using binomial sign test at 90% confidence level (Sheskin 2004)
- Given the results of the LUT 09, the baseline unit would have to lose 4 or more missions to statistically underperform the SUT unit

- Given the starting strength ratio of 2:1, it is unlikely the baseline unit will lose 4 missions
- Modify test structure to use a lower starting strength ratio

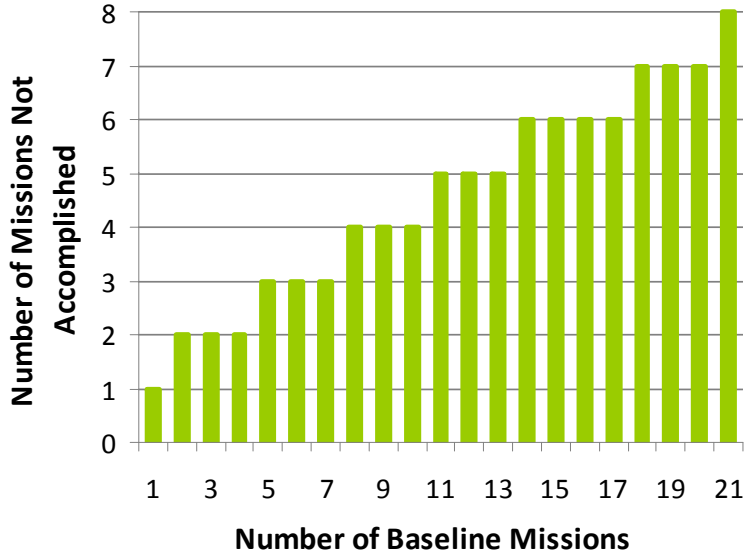
Mission Success Rate

- Comparative evaluation using two proportion z-test at 90% confidence level

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$



- Given an expected 13 baseline missions to be conducted, the required performance of the baseline unit is a maximum mission success rate of 63%



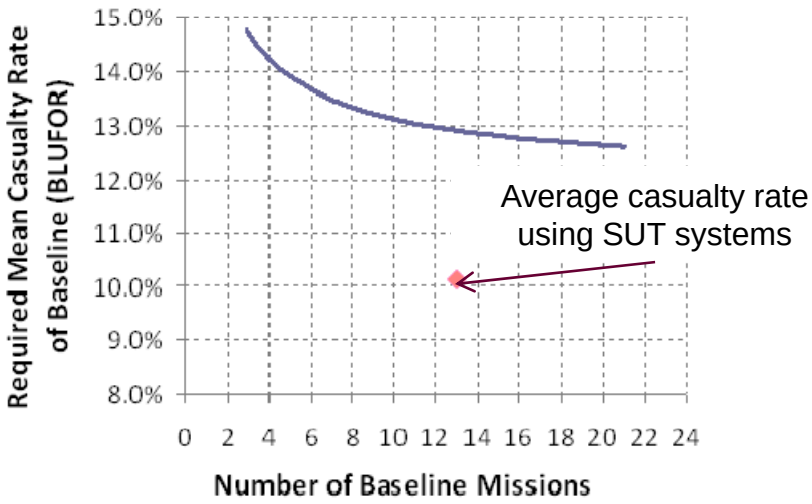
- Given the starting strength ratio of 2:1, it is unlikely that a 63% mission success rate will be observed
- Modify test structure to use a lower starting strength ratio

Casualty Rates

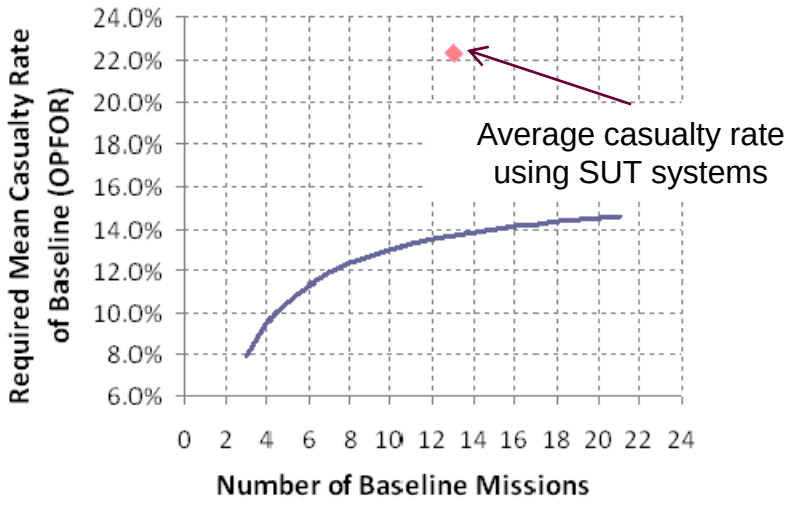
- Comparative evaluation using t-test at 90% confidence level (Sheskin 2004)

$$t_{calc} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Assume variability is the same for both baseline and SUT unit



- Given an expected 13 baseline missions to be conducted:
  - Minimum required BLUFOR rate is 12.9%
  - Maximum required OPFOR rate is 13.7%



- Typical observed BLUFOR and OPFOR rates are around 10% and 25% respectively
- Possible to observe positive impact of SUT on BLUFOR rate, but highly unlikely for OPFOR rate

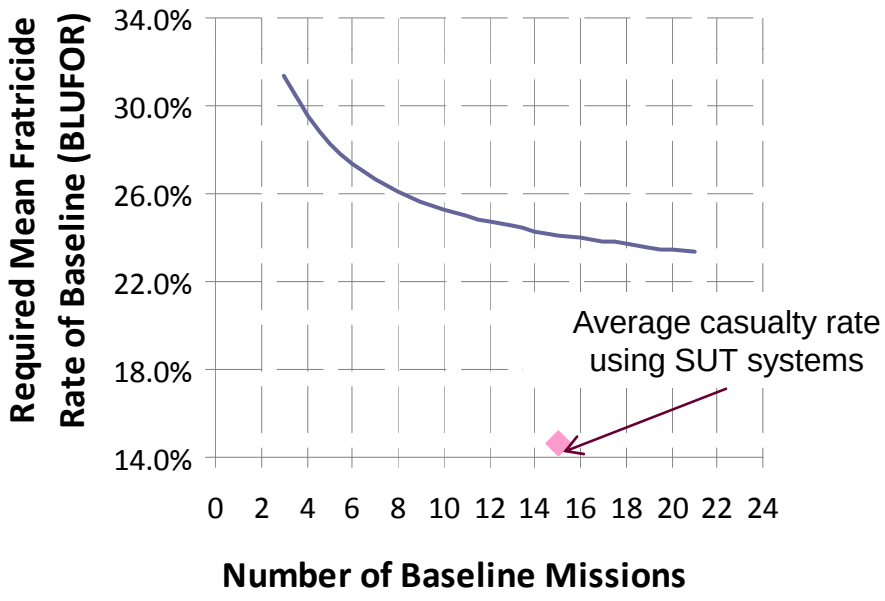
BLUFOR Fratricide Rate

- Comparative evaluation using t-test at 90% confidence level (Sheskin 2004)

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Assume variability is the same for both baseline and SUT unit

- Given an expected 13 baseline missions to be conducted, minimum required BLUFOR fratricide rate is 25%



- Observed BLUFOR fratricides rates are around 13% (Gadsden & Outteridge 2006)
- Highly unlikely to observe significant performance difference between the two units

Sensitivity Analysis

- Analysis predicated on a number of assumptions
  - Variability in performance measures is identical for the SUT and baseline unit
  - 90% confidence interval is the more appropriate confidence interval for the analysis
  - Performance of SUT unit in LUT 09 is representative of future performance in subsequent OT&E

Required casualty rates and mission success metrics are consistent with observed values

Required improved performance of SUT raised concerns about being able to provide conclusive results in a comparative test

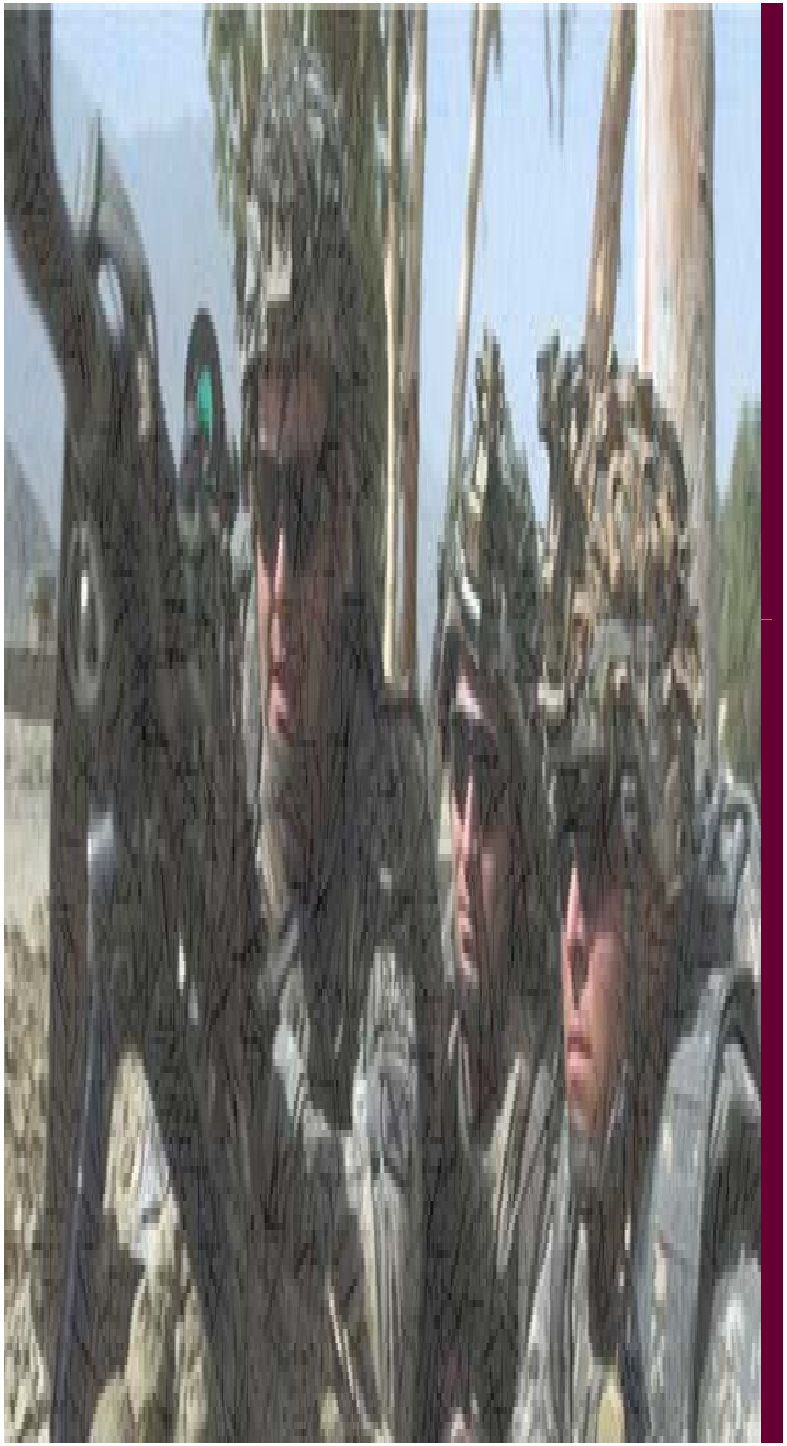
		Required Values for Statistical Significance in IOT&E			
Metrics	**Observed LUT 09**	**Initial Results**	**50% Variability Reduction**	**Confidence Level = 80%**	**SUT**
Missions Not Accomplished	1	4-6	N/A	4	--
Mission Success Rate	0.85	63.2%	N/A	71.1%	98.2%
BLUFOR Casualty Rate	10.1%	12.9%	12.1%	11.9%	4.7%
OPFOR Casualty Rate	22.3%	13.7%	16.2%	16.7%	31.0%
BLUFOR Fratricide Rate	14.6%	24.5%	21.6%	21.2%	7.3%

Required fratricide rate remains high

Required OPFOR casualty rate remains low

Summary

- Using statistical inference insight may be gained about possible outcomes of comparative tests
  - Guide expectations
  - Point to areas where test may need restructuring
  - Enable a more effective utilization of resources
- For case study, it is likely that a comparative evaluation of these quantitative metrics will lead to statistically inconclusive results as performance requirements are high
  - Possible restructuring of test needed
  - Given current performance of SUT, a comparative test may not be an effective utilization of limited resources
- Extend analysis to qualitative measures of operational effectiveness which are gathered from surveys and interviews



Utilizing Statistical Inference to Guide Expectations and Test Structuring during Operational Testing and Evaluation

Joy Brathwaite
School of Aerospace Engineering
Georgia Institute of Technology

Contact: joy.brathwaite@gatech.edu

Dr. Alton Wallace
Dr. Robert Holcomb
Institute for Defense Analyses

May 10-12, 2011

8th Annual Acquisition Research Symposium